Feasibility Study of Composite Dielectric Tunnel Barriers for Flash Memory

Sarves Verma[1], Eric Pop[2], Pawan Kapur[1], Prashant Majhi[2], Krishna Parat[2], Krishna C.Saraswat[1]

1 *Center for Integrated Systems, Stanford University, 20 Via Palou, MS 4075, Stanford, CA 94305*

2. Intel Corporation, Phone: 650-725-3610, email: sarves@stanford.edu

Lack of voltage reduction presents a serious impediment in future flash memory scaling. Replacing conventional SiO_2 tunnel dielectric with a composite dielectric material (combination of high-κ and SiO_2 layers) stack (Fig.1) potentially yields a powerful approach to achieve voltage reduction. The method relies on obtaining non-linearity in gate current vs. gate voltage (V_{gs}) characteristics, such that, low voltages (corresponding to retention and read disturb (V_{read})) render a lower current; whereas, high voltages (corresponding to program and erase) result in a higher current. Related past work involves using a composite stack with a smaller barrier material (high-κ) between two large barrier materials (SiO_2) [1] and a crested barrier approach [2]. However, in these systems, to-date, a top-down approach to optimize the design space for minimum programming voltage (V_{prog}) and equivalent oxide thickness (EOT), while meeting flash retention, read and program-disturb constrains, has not been pursued. Such an approach identifies the best high-κ option by comparisons, yielding lowest possible flash operational voltages.

Towards this end, we explore both the symmetric (low-κ/high-κ/low-κ) and the asymmetric (low-κ/high-κ) composite tunnel barriers using several high-κ materials. There are three possible tunneling mechanisms through a composite barrier depending on the V_{gs} (Fig. 2a). A corresponding simulated current density (J)-V_{gs} (Fig. 2b) indeed demonstrates a higher non-linearity for composite barrier structure compared to the conventional SiO_2 dielectric, thus, yielding a lower V_{prog}. The higher non-linearity stems from a lowering of both the high-κ barrier (modulated by voltage across low-κ) and the increased E-field with V_{gs}. By comparison, for conventional SiO_2 tunnel dielectric, the barrier height is fixed and only one factor- increase in E-field is responsible for the increase in current with V_{gs}.

Five different high-κ materials (HfO_2, La_2O_3, Y_2O_3, ZrO_2, Al_2O_3) were explored using an in-house simulator based on the transfer matrix approach [3] for current calculations. We first choose the high-κ material. Next, we fix the total EOT (high-κ + SiO_2 layers) and vary SiO_2 thickness (T_{ox}). For each T_{ox}, we simulate the J-V_{gs} curves, and obtain the V_{prog} at the required programming current density ($J_{prog} = 3 \times 10^{-2}$ A/cm^2, typical for NAND Flash). By repeating this for different EOTs, we get a family of V_{prog} vs. T_{ox} curves, each exhibiting a minimum V_{prog} (Fig. 3a). Similarly, Fig. 3b shows the voltages corresponding to the maximum allowed retention ($J_{ret} < 2 \times 10^{-16}$ A/cm^2) and read-disturb ($J_{read} < 7 \times 10^{-11}$ A/cm^2) current densities (shown as horizontal dashed lines, also see Fig. 2b), along with the actual voltages encountered during these conditions. In order to meet these constrains, the voltages obtained should be larger (in absolute value) than the horizontal dashed lines. Thus, these constrains manifest themselves in limiting the allowed T_{ox} range for certain EOTs, resulting in a domain down-selection in Fig. 3a (for V_{prog}). Combining the T_{ox} domain down-selection with the V_{prog} vs. T_{ox} plot, we obtain the optimum V_{prog} for each EOT (Fig. 4). The curve also reveals the minimum possible EOT below which no T_{ox} satisfies the Flash constrains. We repeat the process for different high-κ materials to get the lowest possible V_{prog} along with the 1) best material set, 2) the lowest EOT, and 3) the optimum T_{ox} for that EOT (Fig. 4a,b). The effect of adding constrains can be seen in Fig. 5. In general, constrains increase V_{prog} when compared to the minimum obtained in Fig. 3a, except for program disturb (corresponding to $J_{prog} < 7 \times 10^{-6}$ A/cm^2 refer Fig. 2b) which is always satisfied. Further, upon extraction of erase voltage ($J_{erase} > 7 \times 10^{-3}$ A/cm^2) (Fig. 6), the asymmetric stack was found to have a larger V_{erase} than the symmetric one leading to an important conclusion that symmetric stacks are more promising. Finally, Fig. 7 shows the maximum possible operational voltage V_{max} (maximum of V_{prog} and V_{erase} on the floating gate) vs. EOT. The optimization is across all considered high-κ materials. We find that La_2O_3 performs best for a strict read disturb criterion of 3.6 V while HfO_2 outperforms other high-κ materials for a 2.5 V read disturb. It yields the largest operational voltage of ~5–7 V constituting a ~30%-40% voltage reduction over the conventional SiO_2 based Flash cells.

Next, we experimentally corroborate these results using MOS capacitors. A J-V_{gs} comparison between pure SiO_2 tunnel dielectric and an asymmetric tunnel stack of SiO_2/$HfSiON_x$ (~6.4 nm EOT), reveals a higher non-linearity for the composite stack (Fig. 8). A good agreement between simulations and experiments can also be observed in Fig. 8. Further, C-V measurements (Fig. 9) with minimal hysteresis confirmed a high quality composite dielectric stack. Finally, to ensure that tunneling was indeed Fowler-Nordheim (F-N) dominated, F-N slope was calculated (Fig. 10). An excellent linear fit was obtained from which both barrier height and tunneling mass were estimated.

In conclusion, we have developed a novel optimization methodology for deriving the minimum operational voltage of a composite tunnel dielectric stack for Flash memory. The methodology accounts for normal Flash operation constraints: retention, read and program disturb, and reveals that the higher J-V_{gs} non-linearity of these stacks can result in up to 40% voltage reduction. The symmetric stack is found to be more efficient than asymmetric in reducing voltage. The optimum materials along with their thickness are also identified. Experiments confirm the non-linearity in J-V_{gs} using $HfSiON_x$, a high- κ which for the first time is used in the context of Flash memory.

References :[1] Govoreanu et al., *IEEE Electron Dev. Lett.*, vol. 24, pp. 99-101, 2003 [2] K. K. Likharev, *Appl. Phys. Lett.*, vol. 73, pp. 2137-2139, 1998 [3] Y. Ando et al., *J. Appl. Phys.*, vol. 61, pp. 1497-1502, 1987

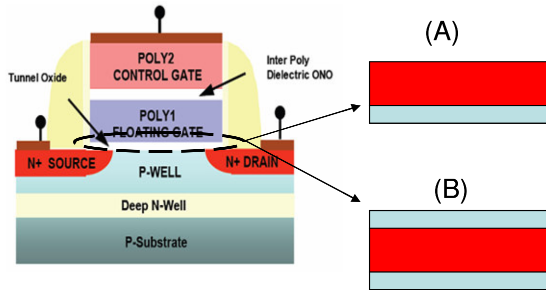

Fig. 1 Schematic of a Flash memory showing the conventional tunnel oxide. (A) Shows replacement of the tunnel stack by an asymmetric barrier while (B) Shows that by a symmetric one. The thicker layer represents high-κ material while thinner represents SiO_2

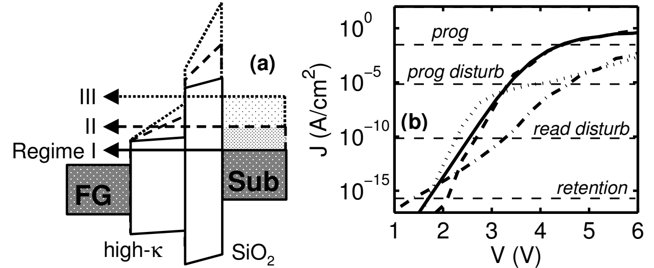

Fig. 2.a Electron substrate injection in Regimes I-III exhibiting different tunneling mechanisms. Fig. 2b Shows the J-V characteristics of asymmetric stack (solid line, T_{ox} = 2 nm), symmetric (dashed, T_{ox} = 2 nm on either side), and pure SiO_2 (dash-dot) with 6 nm total EOT. The dotted line is for a thicker T_{ox} = 3.5 nm in the asymmetric 6 nm EOT stack. The four horizontal dashed lines are Flash constrains mentioned in the text

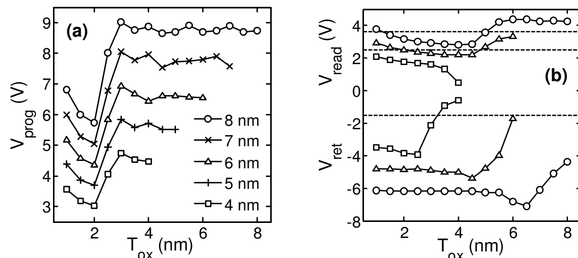

Fig. 3a V_{prog} scaling with T_{ox} for different asymmetric EOTs with HfO_2 (note the minima). Fig. 3b shows the read disturb (top) and retention voltage (bottom) scaling with T_{ox} for the 4, 6 and 8 nm EOT stacks. The top horizontal dashed lines correspond to 2.5 and 3.6 V read disturb voltage; the bottom one corresponds to -1.5V retention. Symbols for different EOTs are consistent across both figures

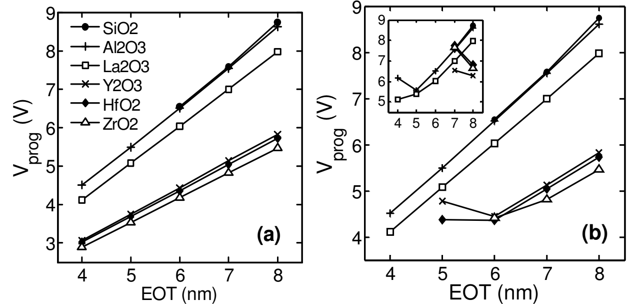

Fig. 4. Optimal V_{prog} at each EOT for all asymmetric stacks and high-κ materials. (a) Imposes only the retention constraint, while (b) adds in the V_{read} = 2.5 V read disturb, and considers the more restrictive 3.6 V in the inset. Symbols are used consistently across the figures.

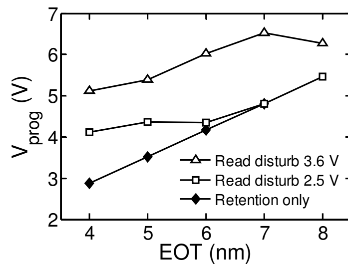

Fig. 5. Change in V_{prog} for asymmetric barriers with different constraints: retention only (no read disturb), and with different read disturb criteria V_{read} = 2.5 and 3.6 V. The values plotted represent global minima, across *all* high-κ materials considered here.

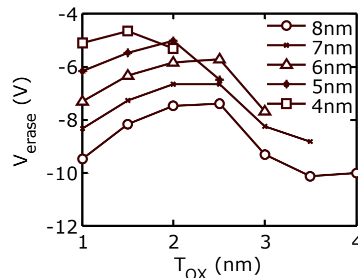

Fig. 6 Erase simulations for a representative case of SiO_2/ HfO_2/SiO_2 symmetric stack for different oxide thicknesses (within the same EOT) and with varying EOT.

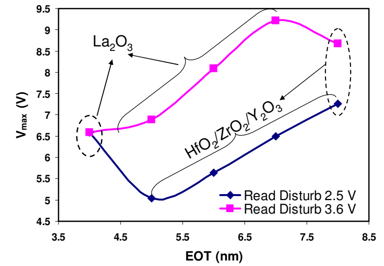

Fig. 7 Global performance optimization for all stacks in consideration irrespective of the high-κ material considered. Further, all constrains like retention, read disturb, program disturb and erase have been taken in account to find the maximum operating voltage (V_{max}) required for a Flash memory. Note only La_2O_3 sustains the stringent read disturb criterion of 3.6 V

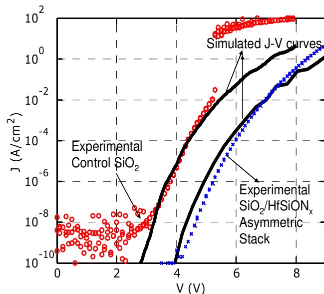

Fig. 8 Experimental J-V curves for a control oxide (SiO_2) of 5.6 nm thickness and for an asymmetric $SiO_2/HfSiON_x$ stack of 6.4 nm EOT. The dotted lines represent experimental results while the solid lines are simulated curves for pure SiO_2 stacks of same EOT. Note that the asymmetric stack has a higher non-linearity when compared to simulated pure SiO_2 of the same EOT. Also note that the simulations do not account for the breakdown in oxides. All samples are made on n-type substrates and for each sample three measurements were taken accounting for different areas.

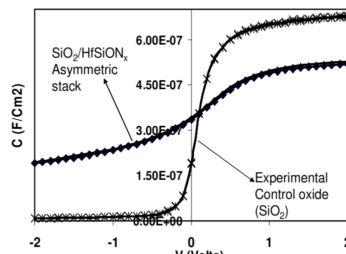

Fig. 9 C-V measurements for the same stack as in Fig. 8.Note that the high-κ /SiO_2 asymmetric stack shows no hysteresis implying a good interface with the Si substrate. Both samples were prepared under different conditions and have different doping levels. For each sample, C-V measurements were taken for 10 KHz, 100 KHz and 1 MHz frequencies.

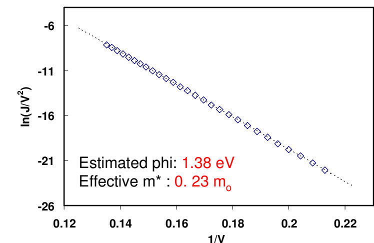

Fig. 10 Plot of Fowler-Nordheim slope for $SiO_2/HfSiON_x$ asymmetric stack (the J-V of which has been shown in Fig.7). Barrier Height (phi) and effective mass (m^*) have hence been derived.